contacts, those farther from the diagonal (parallel or antiparallel to the diagonal) correspond to contacts between β-strands within β-sheets. Thus, the following energetic bias was introduced towards such a mode of packing:

$$E_{map} = -\varepsilon_{gen} \{\Sigma\Sigma (\delta_{ij} \bullet \delta_{i+1\,j+1} \bullet \delta_{i-1\,j-1}) \delta_{par} + \Sigma\Sigma (\delta_{ij} \bullet \delta_{i-1\,j+1} \bullet \delta_{i+1\,j-1}) \delta_{apar}\} \quad (5)$$

where the summations are over all pairs of residues i,j, and δ_{ij} is equal to 1 (0) when residues i and j were (were not) in contact. δ_{par} was equal to 1 only when the corresponding chain fragments are oriented in a parallel manner, *i.e.*, when the chain vectors satisfied the following condition $(v_{i-1} + v_i)\bullet(v_{j-1} + v_j)>0$; otherwise, $\delta_{par} =0$. Similarly, δ_{apar} was equal to 1 when the chain fragments were antiparallel, and it was equal to zero otherwise. For a given contact of a pair of residues, the maximal energetic stabilization due to regular side chain packing was therefore equal to $-\varepsilon_{gen}$, which had the same value as in the previously defined potentials.

The packing cooperativity of the model protein was further enhanced by a term that mimics main chain hydrogen bonds. The geometry of protein hydrogen bonds was translated into a specific range of the model chain geometry. First, a vector was defined that was likely to connect the model beads within motifs that represent regular secondary structure elements. Such a vector should connect beads i and i+3 in a helix and the appropriate beads in a β-sheet. An optimization procedure leads to the following definition of this vector:

$$\mathbf{h}_i = 3.3\ (\mathbf{v}_{i-1} \otimes \mathbf{v}_i)/|(\mathbf{v}_{i-1} \otimes \mathbf{v}_i)| - \mathbf{v}_{i-1}/|\mathbf{v}_{i-1}| \quad (6)$$

The value of the 3.3 pre-factor has been found to be optimal (or more precisely near optimal) for reproducing the internal main chain hydrogen bonding in the lattice projected PDB structures. However, due to the wide distribution of the model chain bond lengths, there were always some hydrogen bonds that were missed in the model. The coordinates of the vectors $\mathbf{h}_i$ were rounded-off to the nearest integer value. Thus, in a helix the $\mathbf{h}_i$ vectors have a component whose length was about 3 lattice units in the direction perpendicular to the three-residue plane (the first

term in the above sum) and were also tilted back by a lattice unit (the last term of equation 6). The projection along the helix axis was also about 3 lattice units; this nicely coincided with the 1.5 Å longitudinal increment per residue in a real helix. Residue i was considered to be hydrogen bonded with residue j when the vector $\mathbf{h}_i$ pointed to any of the 19 points of the excluded volume cluster of residue j. Correspondingly, the vector $-\mathbf{h}_i$ may point to another cluster. Such a situation was illustrated in Figure 13, where residue i is hydrogen bonded with residues j and k because the hydrogen bond vectors coincide with the excluded volume of both residues. The excluded volume clusters were symbolically represented by open spheres. Since the excluded volume clusters never overlapped, the maximum number of these "hydrogen bonds" originating from residue i was equal to 2. The total energy of the "hydrogen bond network" could be written as:

$$E_{H\text{-}bond} = -\varepsilon_{H\text{-}bond} \Sigma(\delta^+ + \delta^- + \delta^{+,-} \quad (7)$$

where δ^- (δ^-) equaled 1 when the vector $\mathbf{h}_i$ ($-\mathbf{h}_i$) connected with an excluded volume cluster, and $\delta^{+,-} = 1$ when the both vectors connected to some clusters, respectively. Otherwise, the corresponding terms were equal to zero. The cooperative contribution, $\delta^{+,-}$, corresponded to local saturation of the hydrogen bond network.

Again, a computational experiment was done to check the effect of these generic potentials on the behavior of the model system. When only the interactions outlined up to this point were included (all the above short- and long-range generic potentials), the model lacked sequence specific information. At sufficiently low temperatures, the chain adopted either of the following two types of structures, a long (sometimes broken) helical structure or a β-sheet with a right-handed supertwist. These motifs fluctuated and were not structurally unique. In a long chain, these two classes of secondary structure elements sometimes formed separate domains.

3. Sequence specific short-range interactions

For the sequence of interest, from the structural database, one may extract the statistics of distances between a pair of amino acids (with their interaction centers as defined in the model) A_i and B_{i-k}, where A and B denote the identities of the amino acids and i is the position in the chain. Here, k=l, 2, 3, 4, 6 and 8 was considered. The terms for k=3 and k=6 were treated as chiral variables. This meant that the distance between A_i and B_{i+3} was stored as a positive or negative number, depending on the handedness of the corresponding three-bond segment. For the k=6 case, the chirality was defined for three subsequent supervectors (the doublet of vectors between beads i and i+2, i+2 and i+4, and from i+4 to i+6). As was done here, the sequence of interest could be divided into overlapping short fragments. These could be aligned to the sequences of known structures. The highest scoring fragments provided a set of structural templates. The obtained statistics could be related to a random distribution and the statistical potential of mean force could be appropriately derived. Terms for k=1, 2, 3, and 4 were weighted equally, while the terms for k=6 and k=8 had weights reduced by a factor of two, with respect to lower order terms. Homologous proteins were always excised from the structural database for the purpose of these test calculations. As previously shown, this type of potential very nicely reproduces the local conformational propensities of globular proteins.[17]

The short-range potentials could be made even more sequence specific when evolutionary information encoded in homologous sequences was employed. In such a case, the aligned fragments of highly homologous sequences (from the sequence database) were treated as the original test sequence, thereby increasing the strength of the statistics. The details of the derivation procedure are given in Appendix 1.

4. Sequence specific pairwise interactions

The pairwise interactions between model residues were defined by contact potentials in the form of a square well function.